# Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM

*Year: 2021* | *Citations: 925* | *Authors: D. Narayanan, M. Shoeybi, J. Casper*

## Abstract

Large language models have led to state-of-the-art accuracies across several tasks. However, training these models efficiently is challenging because: a) GPU memory capacity is limited, making it impossible to fit large models on even a multi-GPU server, and b) the number of compute operations required can result in unrealistically long training times. Consequently, new methods of model parallelism such as tensor and pipeline parallelism have been proposed. Unfortunately, naive usage of these methods leads to scaling issues at thousands of GPUs. In this paper, we show how tensor, pipeline, and data parallelism can be composed to scale to thousands of GPUs. We propose a novel interleaved pipelining schedule that can improve throughput by 10+% with memory footprint comparable to existing approaches. Our approach allows us to perform training iterations on a model with 1 trillion parameters at 502 petaFLOP/s on 3072 GPUs (per-GPU throughput of 52% of theoretical peak).